



# Common Problems in Delay-Based Congestion Control Algorithms: A Gallery of Solutions

M. Rodríguez-Pérez\*, S. Herrería-Alonso, M. Fernández-Veiga, C. López-García

*Dept. Telematics Engineering, E.I. Telecomunicación, Campus Universitario Lagoas-Marcosende s/n, 36310 Vigo, Spain*

## SUMMARY

Although delay-based congestion control protocols such as FAST promise to deliver better performance than traditional TCP Reno, they have not yet been widely incorporated to the Internet. Several factors have contributed to their lack of deployment. Probably, the main contributing factor is that they are not able to compete fairly against loss-based congestion control protocols. In fact, the transmission rate in equilibrium of delay-based approaches is always less than their fair share when they share the network with traditional TCP-Reno derivatives, that employ packet losses as their congestion signal. There are also other performance impairments caused by the sensitivity to errors in the measurement of the congestion signal (queuing delay) that reduce the efficiency and the intra-protocol fairness of the algorithms. In this paper we report, analyze and discuss some recent proposals in the literature to improve the dynamic behavior of delay-based congestion control algorithms, and FAST in particular. Coexistence of sources reacting differently to congestion, identifying congestion appearance in the reverse path and the *persistent congestion* problem are the issues specifically addressed. Copyright © 2011 John Wiley & Sons, Ltd.

## 1. Introduction

Since its early first introduction [1, 2], congestion control has been essential in guaranteeing the stability of the Internet and in enabling its unprecedented growth rate. The original implementation of the control algorithm used packet loss as the triggering signal that sources are aware of to dynamically update the size of their transmission window. More precisely, the absence of an acknowledgement (ACK) or the arrival of several duplicate ACKs are regarded as genuine symptoms of congestion, and the source reacts by shrinking its current window size by half, at least [3]. Over the years, the algorithm has undergone several refinements and modifications in order to control finely the dynamics of end-to-end flows, but the fundamental principle of detecting congestion through packet losses remains untouched in TCP Reno, the most widely deployed TCP version. It is common knowledge, however, that using packet drops to drive the

congestion control actions exhibits some limitations in wireless networks [4] —where discarding packets due to transmission errors is not a rare event— and in networks with a large bandwidth  $\times$  delay product where sharply halving the window size may be an overreaction. This is the motivation behind a plethora of protocol proposals that either propose the use of more feedback information about congestion, like XCP variants [5, 6, 7, 8], or directly advocate the use of congestion signals of a different kind. The class of delay-based congestion avoidance algorithms (DCA) includes, for instance, CARD [9], DUAL [10], Vegas [11], FAST [12] and LEDBAT [13], in chronological order. The premise of DCA schemes is that *queuing delay variation* is a more robust way to detect incipient congestion. Consequently, the senders measure end-to-end queuing delay of data packets and adapt accordingly the transmission rate, an idea that has also been tried for traffic engineering, as pointed out by [14]. There are also recent examples of mixed DCA and loss-based protocols, such as TCP Illinois [15] and compound TCP (CTCP) [16], i.e., solutions using a multiplicity of congestion signals. The

\*Correspondence to: Email address: Miguel.Rodriguez@det.uvigo.es.

latter has been commercially deployed in the Windows 7 and Vista operating systems TCP stacks, although it is turned off by default.

Generally, DCAs present several appealing characteristics to support them as an alternative to the traditional loss-based congestion control protocols. First, in a homogeneous network environment, one where all sources adapt their rates to the same congestion signal (queuing delay), the bandwidth in the bottleneck links is fully utilized, even if *different* source algorithms are used. The reason is that all DCAs carefully avoid packet losses, so the instantaneous transmission rate shows less oscillations than in conventional TCP and TCP-friendly [17] sources. Second, the bandwidth is fairly shared among the competing flows. In fact, DCA sources only react to the average queuing delay, which is common to every flow in a given route,<sup>†</sup> but not on the propagation delay or the buffer sizes. In contrast, recall that, in equilibrium, TCP senders with shorter round-trip times (RTT) attain greater throughput [18, 19]. Third, some DCAs, notably FAST, show fast convergence times and very small variations in the instantaneous rate, thus arising as ideally suited for long-fat pipes. Moreover, with the tools of the network utility maximization theory [12, 20, 21, 22, 23, 24, 25], the equilibrium, fairness and convergence properties of DCAs are now well characterized and can be carefully engineered.

Despite these clear advantages, researches have also discovered that DCAs can give rise to a number of anomalous behaviors leading to inefficient or unfair use of the network resources which, ultimately, have hindered their practical adoption. Probably, the most dangerous problem is the coexistence of DCAs with other congestion control protocols. In a heterogeneous network, one where the flows are responsive to different congestion indications, there may exist multiple equilibrium points, depending on the network parameters (e.g., buffer sizes) and the ordering of the flow arrival times [22, 26, 27, 28]. This means that the bandwidth share is, in general, unpredictable and cannot be controllable. Thus, throughput and fairness may seem chaotic in heterogeneous networks. In many cases, when confronted to TCP Reno, DCAs receive far less bandwidth than expected, as loss-based protocols seize the greater part of the bottleneck queue, see [21, 29, 30] for an account. Later, in this paper, we will describe proposed solutions for controlling inter-protocol fairness between DCA flows and loss based flows. Note that

LEDBAT escapes from this typical behavior, its main design purpose being able to saturate the bottleneck, while yielding to standard TCP. A unique network equilibrium point, optimum in the sense of maximizing a weighted aggregate utility function, can be enforced by imposing on all the flows two conditions, the access to a common price and the adoption of a common slow-timescale adaptation rule [31]. The use of a weighted utility function implies that there is some efficiency loss and inter-protocol unfairness, though.

Another impairment is the *persistent congestion problem* [12, 32, 33], which is a side effect of the procedure employed for detecting congestion. For proper operation, DCAs need to measure queuing delay, and they do this indirectly, usually estimating the round-trip time and the propagation delay (some newer protocols, like LEDBAT, estimate directly the one-way delay (OWD)). The queuing delay is the difference between both, and the propagation delay estimation is simply the minimum of the observed RTTs. Note that these measurements are a purely local procedure. As DCA flows maintain a constant amount of traffic (the precise value is a protocol configuration parameter) queued in the network, newly established flows are very likely to overestimate the propagation delay and assume a false available bandwidth. The consequence is a severe unfairness among the flows. More importantly, this situation tends to persist in time as long as the mean number of active flows does not vary much.

Finally, a third practical limitation of most DCAs comes from their inability to distinguish between congestion in the forward path from congestion in the reverse path. More precisely, when there appears congestion in the return path from receiver to sender, DCAs wrongly assume that the queuing delay has increased, and react diminishing their sending rate [34, 35, 36, 37].<sup>‡</sup> The reduction in efficiency can be large even if the delay of the ACKs grows only moderately, but is especially harmful in paths with large bandwidth  $\times$  delay product.

In this paper we will separately analyze each of the three phenomena and comment several solutions proposed in the literature in the last few years. The material is not new, but it has appeared scattered in a number of papers up to now and may not be well known. We hope that the review can help to clarify the role of DCAs as well as their potential benefits, and ease the way for their deployment as general purpose congestion control mechanisms. Rather than a replacement of current practice in the field of congestion

<sup>†</sup>Clearly, we are not considering the possibility of differentiated services in the network.

<sup>‡</sup>Obviously, protocols that just employ OWD, like LEDBAT, are immune to this problem.

control, congestion based on delay must be considered as a solution to coexist for a long time with the binary-feedback loss-based approach.

The rest of this paper is organized as follows. Section 2 presents a comprehensive description of FAST as a representative of the current paradigm of DCA algorithms. In section 3 we present an analysis of the reverse-path congestion problem and several solutions to it. Section 4 deals with the causes of the persistent congestion problem and explains how to solve it. The inter-protocol fairness problem is discussed in Section 5 and, like with the two previous problems, a solution is described. Finally, the conclusions are summarized in Section 6.

## 2. FAST: throughput, fairness, stability and optimality

DCAs turn out to be better than TCP Reno and its variants for data transmission over large bandwidth  $\times$  delay paths, where packet losses are too scarce to allow the timely adjustment of the sending rate. They also offer better performance to those applications impaired by sudden changes in the transmission rate. Their basic assumption is that it is possible to gain insight into network status observing the variations in the RTT, because the difference between the RTT and the propagation delay is directly related to the amount of data enqueued. So, larger differences imply nearness to congestion. By adjusting the window size based on these variations, DCA-controlled flows are able to keep a high transmission rate without inducing congestion. In comparison, TCP Reno (and other protocols with similar reactions) operate slowly pushing the network to congestion in order to find out the largest possible window size, namely until a packet loss occurs. The current window size is then reduced by a multiplicative factor and the increment phase starts over again.

Since both Vegas and FAST have a similar *modus operandi* (the latter can actually be modelled as a generalized version of the former [12]), and all DCAs display strong similarities among them, we shall describe in this Section a basic model for FAST, taken as a representative of the entire class of algorithms. Most of the remarks in the following can also be applied to other cases with no or only minor adjustments related to the specific utility function of the protocol.

Like TCP Reno, FAST regulates the window size  $w(t)$  in order to adjust the transmission rate. At the flow level, the window size varies dynamically as dictated by the equation

$$\dot{w}(t) = \gamma \alpha \left( 1 - \frac{q(t)x(t)}{\alpha} \right), \quad \gamma \in (0, 1] \text{ and } \alpha > 0 \quad (1)$$

where  $\gamma$  and  $\alpha$  are configuration parameters,  $q(t)$  is the instantaneous queuing delay and  $x(t) = \frac{w(t)}{d+q(t)}$  is the transmission rate, where  $d$  denotes the flow's round trip propagation delay. This equation is just a form of proportional control law, in that the rate of change in the window size is (multiplied by a scale factor) equal to the distance to equilibrium  $x^*q^* = \alpha$ . Hereafter, starred symbols refer to quantities in equilibrium. Proportional control usually yields very fast convergence times to a stable point, when it exists.

The above dynamic flow level behavior is implemented at the packet level with the following rule. Every update interval, defined to be a constant time or some number of RTTs, depending on the specific FAST version, the window size is updated as

$$w_{i+1} = \gamma \left( \frac{\hat{d}w_i}{\hat{r}} + \alpha \right) + (1 - \gamma)w_i, \quad i = 0, 1, 2, \dots \quad (2)$$

where  $\hat{d}$  is the current estimation of the round-trip propagation delay,  $\hat{r} = \hat{d} + \hat{q}$  is an estimation of the RTT, and  $w_0$  is the selected initial window size. The accurate estimation of  $d$  is a bit tricky, as it can only be correctly measured in the absence of interfering or background traffic. In practice,  $\hat{d}$  is set to the minimum RTT observed during the whole transmission. In the end, this is only a problem when different FAST flows have different overestimations. As long as all FAST flows make the same error, the fairness properties are not affected. Later in this paper, we will study this problem and explain some solutions for it. Eq. (2) brings light to the meaning of the constant  $\gamma$ . Actually, it works as a *smoothing factor* or *gain* that controls the speed of convergence. Its value is taken freely from the interval  $(0, 1]$ , although  $\gamma = 0.5$  is a common choice.

The equilibrium properties of FAST are well established in the literature [12, 38] and coincide with those of Vegas [29, 39]. Fixing  $\dot{w}(t) = 0$  in (1), it is immediate to see that, under equilibrium, each flow achieves a throughput

$$x^* = \frac{\alpha}{q^*} = \frac{\alpha}{\hat{r}^* - \hat{d}}. \quad (3)$$

Note first that, if a set of FAST flows are equally configured (they have the same  $\alpha$ ) and share a bottleneck link, they all have the same equilibrium rate  $x_i^* = x^*$  because the queuing delay  $r^* - \hat{d}$  is the same, irrespective of their propagation delays. Secondly, the physical meaning of  $\alpha$  is easily revealed: the product of the queuing delay by the aggregate transmission rates equals the total amount of data

in transit. Assuming that there is a unique bottleneck link along the path with bandwidth  $C$

$$(\hat{r}^* - \hat{d}) \sum_{i=1}^n x_i^* = (\hat{r}^* - \hat{d})C = n\alpha, \quad (4)$$

for  $n$  flows, it follows from (4) that each flow contributes  $\alpha$  packets to the bottleneck queue backlog.

There is a trade-off for selecting the proper value of  $\alpha$ . On the one hand, we would prefer a small value, to minimize overall latency and buffering needs in the network. However, this makes it more difficult to accurately measure the queuing delays, since those will be small too. On the other hand, large values for  $\alpha$  provide faster convergence times.

Considered as a distributed algorithm, the marginal utility function of a FAST source is the queuing delay  $\alpha/x^*$ , whereas it is  $\kappa/(w^*)^\beta$  for the family of loss-based congestion control protocols, with  $\kappa > 0$  an implementation-dependent constant. For instance,  $\beta = 2$  for the classical TCP Reno,  $\beta = 1$  for SCTP [40] and  $\beta = 1.2$  for HighSpeed TCP [41]. The equilibrium point in a homogeneous network of FAST sources with arbitrary topology is unique and optimal, i.e., it optimizes aggregate utility. The fairness properties are immediate from (3) and the global stability of the algorithm has been proved in [25] for the case of homogeneous flows. As expected, not every  $(\alpha, \gamma)$  combination produces stable configurations and several other papers deal with the mathematical conditions to reach a stable equilibrium, see [23, 24, 42]. For conditions about stability and optimality in heterogeneous networks we refer the reader to [31].

### 3. The reverse-path congestion problem

Like TCP Reno, FAST assumes that all the congestion it is measuring happens in the data forward path. However, this assumption is not always true. Congestion could very well occur in the return path, rendering any reaction to this congestion futile. In the last years, the number of asymmetric links installed in the Internet has grown substantially, mainly from residential access lines (xDSL), increasing the likelihood of congestion in the return path. We will show that the onset of reverse path congestion can severely degrade the performance of DCAs. Indeed, the appearance of reverse path congestion has two direct consequences.

Firstly, it causes some acknowledgement packets to get dropped. TCP sources will treat these losses like ordinary data packet losses, namely slowing down their sending

rates. The problem affects all TCP variants and is not just a drawback in FAST. However, it is not particularly harmful unless many consecutive ACKs are dropped, since TCP ACKs are cumulative and the loss of one ACK is repaired by the reception of a subsequent one [34]. The standardized TCP SACK [43] allows selective acknowledgements just for recovering from multiple lost segments within a RTT. The second consequence only affects protocols that directly use the variations of the round-trip time to estimate congestion, like FAST or Vegas. Recall from Section 2 that the equilibrium throughput is inversely proportional to the total queuing delay, cf. (3). If we expand  $r^* = q_f^* + q_b^* + \hat{d}$ , where  $q_f^*$  is the (equilibrium) forward queuing delay and  $q_b^*$  is the (equilibrium) reverse path delay, we get

$$x^* = \frac{\alpha}{q_f^* + q_b^*}. \quad (5)$$

From (5) it is clear that backward queuing delay has as much weight as the forward queuing delay to establish the final operating point, when really only the forward queuing delay should have been taken into account. At last, this is the one that can be reduced after diminishing the transmission rate.

It is possible to quantify the effect of backward queuing delay in the throughput when the propagation delay is  $d$ . Let us make  $d = kq_f^*$  and  $\rho = q_b^*/r^*$ , where  $k$  is a suitable factor. We can solve for  $q_b^*$  to obtain

$$q_b^* = (k+1) \frac{\rho q_f^*}{1-\rho}. \quad (6)$$

Substituting (6) into (5) we find that

$$x^* = \frac{\alpha}{q_f^*} \frac{1-\rho}{1+k\rho}. \quad (7)$$

In Fig. 1 we have represented the decay in the equilibrium throughput of a FAST connection as the backward to total queuing delay increases, for different round-trip propagation delays. Obviously, the best case is when the propagation delay is negligible compared to the queuing delay. Also, the impact on throughput is large for backward queuing delays greater than around 10 % of the total delay.

#### 3.1. Network-assisted solution

Solutions to the significant fall in performance due to the contribution of reverse path delay came in two ways, network-assisted and pure end-to-end approaches.

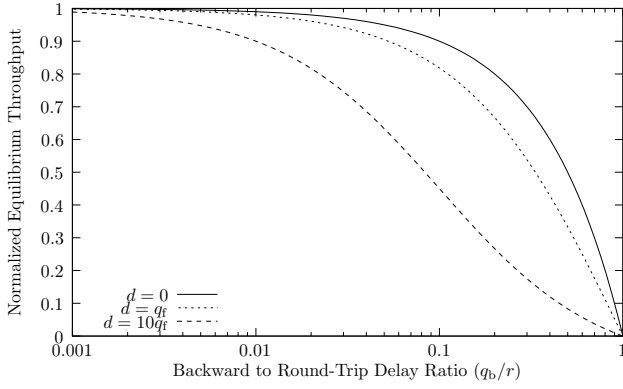


Figure 1. Effect of the backward delay queuing  $q_b$  on FAST throughput for different values of the propagation delay  $d$  and a fixed forward queuing delay  $q_f^* > 0$ .

The most useful technique of the first type is probably [36]. At least, it requires the smallest number of changes to the existing network architecture. Though designed with Vegas in mind, it may be used without modification in FAST. The authors argue that  $r^* - d$  is not a correct way to measure queuing delay, since that difference counts the return path queuing delay too. But, if the delay in the reverse path  $r_b^*$  could be measured separately, subtracting it from  $r^*$ , a more accurate estimation  $r' = r^* - r_b^*$  allows the source to adapt its rate correctly.

The idea used in [36] to calculate  $r_b^*$  is to install modified RED [44] queues in the return path that mark the packets with probability

$$p_{ECN} = \frac{(b/C - \min_{th})^+}{\max_{th} - \min_{th}}, \quad (8)$$

where  $b$  is the average queue length,  $C$  the link bandwidth and  $\max_{th}$ ,  $\min_{th}$  the usual RED thresholds. Now, the change in  $p_{ECN}$  is found to be *proportional* to the change of  $r_b^*$ , so tracking  $p_{ECN}$  is a way of inferring how much delay contributes the reverse path.

Despite its simplicity, the method has some shortcomings. One is that it depends on the deployment of modified RED queues on the return path. The second, and more serious, is that it fails when RED queues are also used in the forward path. Any other router that also did mark packets would alter the amount of change in  $r_b^*$  calculation, effectively making its value useless. A last drawback is that the new estimation  $r'$  does not eliminate the return path congestion problem in every possible scenario.

### 3.2. End-to-end solutions

Enhanced Vegas [35] and LEDBAT [13] present alternative proposals to correct the overestimation of the round-trip time due to the return path congestion. Unlike other approaches, they are pure end-to-end solutions requiring no explicit support in the network. Both Enhanced Vegas and LEDBAT exploit the TCP timestamp options to accomplish their goal. Their use for LEDBAT is trivial, as it only requires to measure changes in one-way delay to infer congestion. So, simply adding a timestamp to the data segments and a measurement result field in the ack packets is sufficient for it.

However both FAST and Vegas employ the queuing delay, and not just its variations, so the LEDBAT method can not be used. Enhanced Vegas uses the TCP timestamp options so as to compute an accurate estimation of the backward queuing delay of the ACK segments. Specifically, the measurement algorithm estimates the backward trip time  $r_b$  and then deduces the backward queuing delay from that value. Synchronizing the system clocks of the TCP sender and receiver is not a requirement for sampling  $r_b$ , but unfortunately aligning both clock speeds is necessary. Otherwise, if the clocks drift, the measured values become erroneous. The problem may be solved by resorting to an external (i.e., not part of the transport entities) procedure to quantify the clock skew itself and compensate for its value [45, 46]. Provided this is the case, the backward trip time is then subtracted from the total RTT  $r' = r - r_b$  exactly as in [36], so that it only includes the term for the forward queuing delay and the round-trip propagation delay.

While suppressing the backward delay removes the effect of congested reverse paths, or similarly that of delayed ACKs or asymmetric forward and backward routes, in most cases, it is not a complete fix, as reported in [37] whose analysis is briefly reproduced here.

Recall (2), the congestion window update in FAST. Consider for simplicity, but without loss of generality, the case  $\gamma = 1$ . In equilibrium it must hold that

$$w^* = \frac{\hat{d}w^*}{\hat{r}'} + \alpha, \quad (9)$$

where  $\hat{r}' = \hat{r}^* - \hat{q}_b^* = \hat{q}_f^* + \hat{d}$ . Reordering the terms, we have

$$w^* \left( 1 - \frac{\hat{d}}{\hat{r}'} \right) = \alpha, \quad (10)$$

but  $1 - \hat{d}/\hat{r}' = q_f^*/\hat{r}'$ . Using this identity and (10) we obtain the equilibrium value of the window size as

$$w^* = \alpha \frac{\hat{r}'}{\hat{q}_f^*}. \quad (11)$$

Finally, from (11) and the well-known relation  $w^* = x^* r^*$ ,

$$x^* = \frac{\alpha}{\hat{q}_f^*} \left( 1 - \frac{\hat{q}_b^*}{\hat{r}^*} \right). \quad (12)$$

Consequently, the term  $1 - \hat{q}_b^*/\hat{r}^*$  still introduces a bias on the throughput as the backward congestion increases. This situation corresponds exactly with the case  $d = 0$  depicted previously in Fig. 1. So, either the network-assisted or the end-to-end solutions discussed above might help to improve throughput in case of backward congestion, but are effective only if the accumulated queuing delay in the return path is a small fraction of the total round-trip time.

A simple fix for removing completely the deviation caused by  $q_b^*$  was presented in [37]. Let  $r^*$  stay as is and modify  $d'$  as  $d' = \hat{d} + \hat{q}_b^*$ . It is elementary to plug  $d'$  and  $r^*$  in (9) and find that now

$$x^* = \frac{\alpha}{\hat{q}_f^*} \quad (13)$$

which is the exact equilibrium rate. In conclusion, *DCA's can be made robust against the delay in the reverse path only if the queuing delay of ACKs is regarded as part of the propagation delay.*

#### 4. The persistent congestion problem

DCA's are very sensitive to measurement errors in the round-trip time and the propagation delay. FAST is no exception to this rule. The problem lies in the way the queuing delay is computed. While the RTT estimation  $\hat{r}$  can be easily computed by the sender, for each data segment, the value of the propagation delay is not directly available and has to be inferred somehow. Most DCA protocols use the same simple heuristic to get  $d$ , consisting on identifying  $\hat{d}$  to the minimum  $r$  observed throughout a connection lifetime. This works correctly as long as the intermediate buffers eventually empty, at all the routers traversed by the packets, and the network path they follow does not change. In any other case, the propagation delay is incorrectly measured, and the difference  $\hat{r} - \hat{d}$  does not truly represent the queuing delay.

If the network changes the route between sender and receiver, and particularly when the new path is longer,

DCA-controlled flows overestimate the queuing delay because they record an outdated estimation of  $d$ . However, it is fair to say that rerouting does not represent much of a concern in practice, and that there exist simple ways that most DCA implementations can follow to ignore the issue. For instance,  $\hat{d}$  can be periodically replaced with the minimum observed  $r$  in the last monitored period. This, at least, limits the duration of the effects due to rerouting.

On the other hand, ensuring that network buffers get eventually empty and there are chances for a DCA flow to find out  $d$  is more difficult.<sup>§</sup> Every time a connection is unable to get the real propagation delay, because intermediate buffers have some backlogged data, the queuing delay is underestimated. As a result, the network path falls into a state of *persistent congestion*, which is the direct cause of two different problems.

For instance, taking FAST as an example, overestimating the propagation delay makes FAST flows try to buffer more data in the network than they are allowed to do. Let  $\hat{d} = k d$ , with  $k > 1$ , where  $d$  is the real round-trip propagation delay and  $\hat{d}$  is the measured one, and consider for simplicity a single FAST flow offering traffic to a bottleneck link with capacity  $C$ . From (3) we derive

$$\alpha = C (r^* - \hat{d}). \quad (14)$$

Because  $k > 1$ ,  $r^* - \hat{d} < r^* - d$ , and since the queue length is  $l = C(r^* - d)$  it is easy to see that  $l > C(r^* - \hat{d}) = \alpha$ .

The second problem caused by persistent congestion is worse, in that it leads to intra-protocol unfairness. Consider  $n$  FAST flows arriving consecutively to a bottleneck link. It is fairly obvious that the latest flow to arrive at the bottleneck will overestimate its propagation delay, as it will have to account for the traffic queued by the previous flows. As shown in [33], in that case the ratio of throughput between the last and the first flows is, in equilibrium, at worst,  $O(n)$ .

Although it is arguably not the common pattern that flows arrive consecutively with the older flows never leaving, the above equation helps to understand the magnitude of the relative unfairness caused by persistent congestion when a handful of flows arrive to a shared bottleneck. We will describe next how the unfairness can be solved.

<sup>§</sup>This possibility should not be overlooked. Consider, for example, a homogeneous network with several long-lasting FAST flows. In such a scenario, the buffers will not empty, since the flows do always maintain  $\alpha$  packets queued in the network.

#### 4.1. Solving persistent congestion with active queue management

Similarly to the case of the return path congestion, solutions to persistent congestion fall in two classes: network-assisted, usually in the form of active queue management algorithms (AQM), and pure end-to-end approaches. We will start analyzing AQM proposals first.

In [47] routers with RED are proposed in order to allocate bandwidth more evenly among the flows, regardless of their starting times. However, finding the appropriate threshold values for the RED gateways is not easy and remains an open issue [48]. Also, this approach does not tackle the root problem, the incorrect estimation of the propagation delay by the late-coming flows, but merely address its consequences. Conceptually, RED breaks the basic assumption of DCAs that packets will not be dropped, since congestion will not build up. Contrarily, RED drops packets before congestion. In the end, one could argue that mixing RED and DCA is useless, because one of the purposes in RED is to warn loss-based congestion controlled sources early by randomly discarding some packets. DCAs are oblivious to these indications.

The same argument applies to [29], where the authors suggest a way to eliminate persistent congestion using REM at the routers. REM [49] is an active queue management scheme that keeps buffer low while sustaining high link utilization. Certainly, with small queues, the minimum of all measured RTTs is a good approximation to propagation delay, but the problem now is that there is not enough information in the queuing delay to enable the detection of congestion. In fact, the price information with REM is only carried in packet losses, a signal that pure DCAs dismiss. The necessary modifications in the window adjustment policy of Vegas so that it interacts adequately with REM are also presented in [29]. Overall, one could consider this as a mixed DCA- and loss-based protocol, more in the spirit of CTCP.

In [50] a new IP option called AQT (Accumulate Queuing Time) is defined. It is used to collect the queuing time experienced by FAST packets along a path. With this scheme, FAST sources must send some probing packets with the AQT option active, and the routers must compute the queuing time for each received probing packet and add it to the actual AQT field. As a result, each connection is able to obtain a good estimate of the propagation delay, sorting out the queuing time from the RTT measurement. However, the disadvantages (and complexities) are clear: cooperation from all the routers, explicit information exchange and modifications of the IP header.

A similar proposal is [26], which solves the persistent congestion problem by marking the ToS field in the IP header with the highest priority for the first packet of each flow. With priority queuing at the routers, the highest priority packets will be dispatched immediately even if the router buffer is not empty and, therefore, FAST will obtain an accurate estimate of the propagation delay. Despite being simpler than AQT (the ToS field is standardized), it forces the routers to use priority scheduling or recognize the packets from FAST flows.

#### 4.2. End-to-end solutions to persistent congestion

The need for end-to-end solutions to the persistent congestion problem is debatable. On the one hand, users can not trust that adequate AQM schemes are to be widely installed throughout all the Internet as an aid to the use of DCAs. On the other hand, bottlenecks on the Internet are not very likely to be monopolized only by DCA flows, instead of containing a mix of both DCA and TCP Reno data flows. In this latter case, the probability of persistent congestion is low, given that the bottleneck queues will eventually empty. It is due to the inherent dynamics of TCP Reno, and it means that DCA-controlled flows will be able to read the correct round-trip propagation delay. In any case, a source can find a bottleneck link occupied only by DCA flows, the likeliness increasing with the proximity of this router to a DCA traffic sender. There are end-to-end solutions to the persistent congestion problem in those cases too.

It is also questionable whether a given DCA flow has any interest in helping to make persistent congestion vanish. After all, late-coming flows greedily benefit from it, attaining a higher throughput than older flows. We argue, however, that it is in the interest of flows to collaborate in the avoidance of persistent congestion. First, persistent congestion causes larger queues at the bottleneck links, thus larger delays. Secondly, a flow is favoured by the persistent congestion state as long as it is the latest to become active. This is a condition that cannot last forever, so a source is likely to take some actions to counteract the persistent congestion. This gives other flows an incentive to behave similarly.

The key idea of the end-to-end solutions to the persistent congestion problem is giving a chance to *new connections* so that the true round-trip propagation delay can be sampled, without perturbing the flows already

established.<sup>¶</sup> Hence, consider as a starting point an ideal situation where  $n$  flows have measured the propagation delay and look for a method whereby a new flow is able to pinpoint the propagation delay.

The first proposal of such a method appeared in [51]. Here, the key idea is to let the new flow pause its transmission sometime after reaching equilibrium, so as to let the bottleneck queue drain its backlog. It is assumed that, if the pause is long enough, when the transmission is resumed the first packet will find an empty queue and will see directly the path's propagation delay, hence eliminating the persistent congestion bias. However, the pause length must have an upper bound. The queue only drains until the rest of the competing flows discover that there is room for more packets, and then increase their transmission rates. So the queue has to drain completely in at most one RTT, reducing the generality of this approach. It has been found [52] that there is a lower bound to the propagation delay below which the pause is not effective. For a simplified scenario in which all existing flows have the same propagation delay  $d$ , the minimum  $d$  needed to empty the queue in just one RTT is

$$d > \frac{n\alpha\sqrt{1+4n}}{2C} = O\left(n^{\frac{3}{2}}\right), \quad (15)$$

where  $C$  denotes the bottleneck link's bandwidth. Given the scaling with the number of flows, pausing transiently would only work in networks with large propagation delays and a small number of flows.

To overcome these problems, [52] proposes a different behavior for the new flows. The goal now is obtaining the *error* in the estimation of the round-trip propagation delay. This error is just the queuing delay due to the amount of traffic already queued by all the older flows at the bottleneck. If those older flows all are configured with the same  $\alpha$  parameter, which they should if fairness is desired, this delay amounts to  $\epsilon = \frac{\alpha n}{C}$ . The method in [52] estimates indirectly both  $\hat{n}$  and  $\hat{C}$  to evaluate this error (note that  $\alpha$  is already known). For this it takes advantage of a direct relation between the change in transmission rate, the measured change in the round-trip time and  $n$  if two conditions hold [33]: i) the variation lasts a time short enough so that the rest of the flows keep their transmission rates unchanged; ii) the bottleneck queue does not empty during this time. So, provided the two are met, the last arriving flow can get knowledge of  $\hat{n}$ . Once  $\hat{n}$  is known,

<sup>¶</sup>Another possibility would be that old flows detect the arrival of a new flow and react somehow to let the new one seek the right propagation delay. As far as we know, this possibility has not been explored, however.

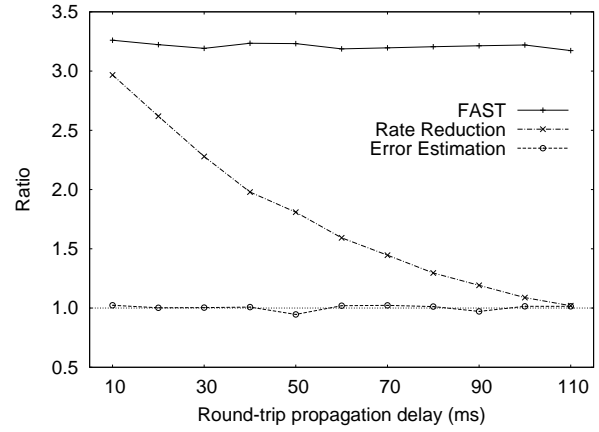


Figure 2. Last arriving flow throughput to previous ones average throughput ratio for different persistent congestion avoidance methods with  $n = 8$ . Taken from [52].

the flow can obtain  $\hat{C}$  using (3) and the relation between  $n$  and the queue backlog under persistent congestion, that is also known [33]. Finally, when both  $\hat{n}$  and  $\hat{C}$  are known, the new round-trip propagation delay  $d'$  can be set to

$$d' = \hat{d} - \frac{\alpha \hat{n}}{\hat{C}}. \quad (16)$$

The key issue is how to modify the transmission rate. If it increases, there is a risk to cause packet drops at the bottleneck, which would make the estimation meaningless. Fortunately, it is very easy to detect such a case. But if the transmission rate is lowered, the queue can empty all its backlog and, as discussed before, the measurement fails too. As this second possibility is much more difficult to ascertain, the authors suggest to modify the transmission rate only by a small increment. In most situations, there should be enough room in the bottleneck buffer to hold some more data for just an RTT, and the failure to do so is not catastrophic.

Just to illustrate the effectiveness of the above solutions, Fig. 2 compares the fairness of original FAST against the solution in [51], labeled in the figure as *Rate Reduction*, and the third approach, labeled *Error Estimation* (EE), for a network with  $n = 8$  flows and different propagation delays. It is clear from the plot that the EE method works no matter what the real propagation delay is. FAST shows a strong deviation against old connections, and the rate reduction approach only works for large propagation delays. Had more flows been set up, fairness among them would have been achieved only for larger propagation delays  $\sim 200$  ms.



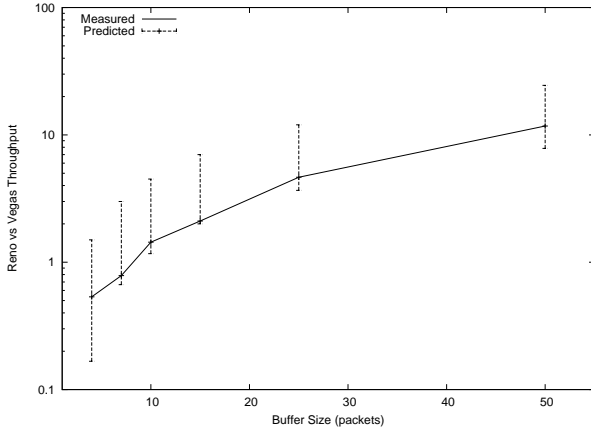


Figure 3. Inter-protocol fairness between TCP-Reno and TCP-Vegas. Original data from [53].

### 5. The inter-protocol fairness problem or the case of FAST against Reno

Probably, the best well known problem of DCAs is that they are generally unable to share the network fairly against the usual TCP variants. This is also the major problem that prevents the wide-scale deployment of general purpose DCA algorithms in the Internet, as other advantages, like fewer packet losses and lower jitter, do not usually pay off receiving less bandwidth. The description of this inter-protocol unfairness problem is found in several papers [30, 32, 53, 54], although usually only in an empirical fashion.

A first informal explanation of the root causes of this unfairness appears in [53], referred to Vegas. The translation to FAST is straightforward, however. Consider the ideal long-run behavior of both FAST and loss-based protocols, in which the former strive to maintain just  $k = \alpha$  packets (or  $\alpha \leq k \leq \beta$  in Vegas), while the latter utilize the full bottleneck size  $B$ . The immediate conclusion is that, at any given time, a FAST flow would have  $k$  packets in the bottleneck, while a TCP Reno flow, as a gross simplification, will have a number in the interval  $[0, B - k]$  or, in average,  $(B - k)/2$  packets. So, for two flows sharing a bottleneck, one running FAST and the other using a Reno-like protocol, their relative bandwidth share should be

$$\frac{x_{\text{Reno}}^*}{x_{\text{FAST}}^*} = \frac{B - k}{2k}. \quad (17)$$

Notwithstanding its simplicity, the above formula provides acceptable approximations to the actual performance of TCP Reno when contending with a DCA. Fig. 3 shows simulations results from [53] where a Vegas flow with

$\alpha = 1$  and  $\beta = 3$  shares a bottleneck of variable buffer size with a TCP Reno flow. Since the exact value of  $k$  cannot be predicted with Vegas, the figure shows both the measured results and the bound (17) with  $k = \alpha$  and  $k = \beta$ . The results also show that the inter-protocol fairness of FAST versus TCP Reno depends at least on the buffer size and the parameter  $\alpha$ . As expected, larger buffer sizes give more advantage to TCP Reno flows, which react more aggressively upon a surplus of bandwidth.

A deeper account of the problem appears in [22]. In this work the authors employ a flow model of both TCP Reno and FAST to study the behavior of a general network where links are potentially shared by both kinds of flows. The problem is formulated as the maximization of the aggregate utility functions of the sources. As the marginal utilities of FAST and TCP Reno are different, and they react to different congestion measures (modeled as link prices), they can not converge to the same operating point. This is expected, as it is in full accordance with (17). Given that the congestion measures, or link prices, are not the same for FAST and TCP Reno, in [22] the authors introduce a *price mapping function* to feed the same link prices to all sources. This allows to reach two important conclusions: i) the intra-protocol fairness properties of TCP Reno and FAST are not affected by the presence of different congestion control protocols; ii) the inter-protocol fairness properties can be adjusted by *simply* multiplying the corresponding utility functions by a constant factor  $\mu$ . This clearly opens the door to new Internet congestion control, as no network changes are needed, not even changes in current congestion control protocols. Only newly deployed protocols have to be adjusted, by a constant factor, to be fair against any other in use.

In the case of FAST, whose utility function is

$$U(x) = \alpha \log(x), \quad (18)$$

the modification needed to make it compatible against TCP Reno is trivial. Multiplying  $U(x)$  by a constant factor is directly equivalent to use a new  $\alpha' = \mu\alpha$ . The difficulty lies in obtaining this new  $\alpha'$  value in a scalable and decentralized fashion.

A first attempt was presented in [27]. The long-term average value of the loss rate is used here to drive the adaption of  $\alpha$  to a value such that FAST competes fairly against TCP Reno. It is then proved that the equilibrium value of  $\alpha$  is given by

$$\alpha^* = \frac{q^*}{\lambda^*}, \quad (19)$$

where  $q^*$  is the queuing delay at equilibrium and  $\lambda^*$  the corresponding loss rate. It is important to note that the above equation implies that in a homogeneous network with DCA-controlled flows and with well dimensioned buffers,  $\alpha^* \rightarrow \infty$  by the lack of packet losses. The method includes a guard against this risk:  $\alpha$  is not modified at all if there are no packet losses.

Clearly, there will be cases where  $\alpha$  is increased from its original value to attain a fair share of bandwidth against TCP Reno. We could be tempted to conclude that this means sacrificing the low latency properties of FAST, as a larger  $\alpha$  means more data queued in the network. However, this is not the case. The total latency remains unaffected, because the larger  $\alpha$  is a consequence of the presence of TCP Reno flows, so the bottleneck queue was already full before  $\alpha$  was increased. Thus, the only effect is that the buffer share used by each flow at the bottlenecks is modified, but not its total size. The main penalty of the algorithm is that it only works in a slow timescale, providing solution to stable scenarios with long-lived FAST flows and a stable amount of TCP Reno flows. More work is needed to devise new end-to-end algorithms than can update  $\alpha$  in smaller timescales.

Another new approach for improving the coexistence between loss based and DCA algorithms is presented in [55, 56]. The respective authors propose methods for the DCA flows to react differently to queuing delay when there are loss-based flows in the network. For this they define two operating scenarios: one when the queuing delay is below a certain threshold and another one when the queuing delay is higher. In the first scenario, with low queuing delay, they behave like normal DCA flows, trying to maintain the queuing delay low. However, the more the queuing delay surpasses the threshold, the less they behave like DCA flows as they begin to progressively ignore the queuing delay feedback. This lets DCA flows compete fairly against loss-based flows. The loss-based flows drive the queuing delay high and DCA flows progressively revert also to loss-based flows. When the loss-based flows abandon the network, the residual reaction to queuing delay drive the operating scenario progressively to lower queuing delay points. The lower the queuing delay, the more the DCA flows respond to queuing delay, finally putting the operating scenario below the threshold where they behave like normal DCA flows.

## 6. Conclusions

Although delay based congestion algorithms, like those present in Vegas or FAST have long promised better

performance than TCP Reno, they have not been deployed in the Internet for several reasons: overreaction to congestion in the return path, intra-protocol fairness problems in networks with persistent congestion and inter-protocol fairness among heterogeneous flows.

In this paper we have reviewed state-of-the-art solutions to all these problems, some requiring changes in the network routers and others consisting purely in end-to-end approaches. We have presented, at least, one end-to-end solution to each of the aforementioned problems.

Our overall consideration is that improved versions of DCAs are ready to be used in the Internet and coexist with the classic loss-based congestion control algorithms. For instance, the fix to the persistent congestion problem produces some variations in the sending rate of the aggregate FAST traffic, but these variations are much smaller than typical rate variations of TCP Reno traffic. Also, the fix for the inter-protocol fairness increases the amount of traffic that FAST sources queue at routers, but only when confronted with TCP Reno and recall that even in that case the end-to-end delay does not get larger because the total queue length does not change.

## ACKNOWLEDGEMENTS

This work was supported by the “Ministerio de Ciencia e Innovación” through the project TEC2009-12135 of the “Plan Nacional de I+D+I” (partially financed with FEDER funds).

## References

1. Jacobson V. Congestion avoidance and control. *Proc. ACM SIGCOMM*, 1996.
2. Allman M, Paxson V, Stevens R. TCP Congestion Control. RFC 2581 apr 1991.
3. Allman M, Paxson V, Blanton E. TCP Congestion Control. RFC 5681 sep 2009.
4. Lassila P, Kuusela P. Performance of TCP on low-bandwidth wireless links with delay spikes. *European Transactions on Telecommunications* oct 2008; **19**(6):653–667.
5. Katabi D, Handley M, Rohrs C. Congestion control for high bandwidth-delay product networks. *SIGCOMM Comput. Commun. Rev.* 2002; **32**(4):89–102, doi:10.1145/964725.633035.
6. Xia Y, Subramanian L, Stoica I, Kalyanaraman S. One more bit is enough. *Proceedings of the ACM SIGCOMM*, Philadelphia, PA (USA), 2005; 37–48.
7. Wu H, Ren F, Mu D, Gong X. An efficient and fair explicit congestion control protocol for high bandwidth-delay product networks. *Computer Communications* 2009; **32**:1138–1147.
8. Almeida A, Belo C. Explicit congestion control based on 1-bit probabilistic marking. *Computer Communications* 2010; **33**:S30–S40.
9. Jain R. A delay-based approach for congestion avoidance in inter-connected heterogeneous computer networks. *SIGCOMM Comput. Commun. Rev.* Oct 1989; **19**(5):56–71, doi:10.1145/74681.74686.

*Euro. Trans. Telecomm.* **22**: 1–11 (2011)  
DOI: 10.1002/ett

10. Wang Z, Crowcroft J. Eliminating periodic packet losses in the 4.3-Tahoe BSD TCP congestion control algorithm. *ACM Computer Communications Review* 1992; .
11. Brakmo LS, O'Malley SW, Peterson LL. TCP Vegas: New techniques for congestion detection and avoidance. *SIGCOMM Comput. Commun. Rev.* 1994; **24**(4):24–35, doi:10.1145/190809.190317.
12. Wei DX, Jin C, Low SH, Hegde S. FAST TCP: Motivation, architecture, algorithms, performance. *IEEE/ACM Transactions on Networking* Dec 2006; **14**(6):1246–1259.
13. Shalunov S, Hazel G, Iyengar J. Low extra delay background transport (LEDBAT). Internet Draft Oct 2010. URL <https://tools.ietf.org/id/draft-ietf-ledbat-congestion-control-03>.
14. Alparslan O, Akar N, Karasan E. TCP flow aware adaptive path switching in Diffserv enabled MPLS networks. *European Transactions on Telecommunications* 2011; In press.
15. Liu S, Basar T, Srikant R. A loss and delay-based congestion control algorithm for high-speed networks. *Proc. First Intl. conf. on Performance Evaluation Methodology Tools (VALUETOOLS'06)*, 2006.
16. Tan K, Song J, Zhang Q, Sridharan M. A compound TCP approach for high-speed and long distance networks. *Proceedings of the IEEE INFOCOM*, Barcelona, Spain, 2006; 1–12.
17. Floyd S, Handley M, Padhye J, Widmer J. Equation-based congestion control for unicast applications. *SIGCOMM Comput. Commun. Rev.* 2000; **30**(4):43–56, doi:10.1145/347057.347397.
18. Padhye J, Firoiu V, Towsley D, Kurose J. Modeling TCP throughput: A simple model and its empirical validation". *SIGCOMM Comput. Commun. Rev.* 1998; **28**(4):303–314, doi:10.1145/285243.285291.
19. Shakkotai S, Kumar A, Karnik A, Avenkar A. TCP performance over end-to-end rate control and stochastic available capacity. *IEEE/ACM Transactions on Networking* Aug 2001; **9**(4):377–391.
20. Low SH, Peterson L, Wang L. Understanding TCP Vegas: a duality model. *SIGMETRICS Perform. Eval. Rev.* 2001; **29**(1):226–235.
21. Kunniyur S, Srikant R. End-to-end congestion control: utility functions, random losses and ECN marks. *IEEE/ACM Transactions on Networking* Oct 2003; **11**(5):689–702.
22. Tang A, Wang J, Hegde S, Low SH. Equilibrium and fairness of networks shared by TCP Reno and Vegas/FAST. *Telecommunication Systems* Dec 2005; **30**(4):417–439.
23. Wang J, Wei DX, H Low S. Modelling and stability of FAST TCP. *Proceedings of the IEEE INFOCOM*, vol. 2, Pasadena, CA, USA, 2005; 938–948.
24. Choi JY, Koo K, Lee JS, Low SH. Global stability of FAST TCP in single-link single-source network. *44th IEEE Conference on Decision and Control*, Seville, Spain, 2005; 1837–1841.
25. Choi JY, Koo K, Wei DX, Lee JS, Low SH. Global exponential stability of FAST TCP. *45th IEEE Conference on Decision and Control*, San Diego, CA, USA, 2006; 639–643.
26. Tan L, Yuan C, Zukerman M. FAST TCP: Fairness and queuing issues. *IEEE Communications Letters* Aug 2005; **9**(8):762–764.
27. Tang A, Wei D, Low SH. Heterogeneous congestion control: Efficiency, fairness and design. *IEEE International Conference on Network Protocols*, Santa Barbara, CA, USA, 2006; 127–136.
28. Tang A, Wang J, Low S, Chiang M. Equilibrium of heterogeneous congestion control: existence and uniqueness. *IEEE/ACM Transactions on Networking* Aug 2007; **15**(4):824–837.
29. Low SH, Peterson L, Wang L. Understanding Vegas: a duality model. *J. ACM* Mar 2002; **49**(2):207–235.
30. Bonal T. Comparison of TCP Reno and TCP Vegas: Efficiency and fairness. *Performance Evaluation* Aug 1999; **36–37**:307–332.
31. Tang A, Wei X, Low S, Chiang M. Heterogeneous congestion control: optimality and stability. *IEEE/ACM Transactions on Networking* Jun 2010; **18**(3):844–857.
32. Hengartner U, Bolliger J, Gross T. TCP Vegas revisited. *Proceedings of the IEEE INFOCOM*, vol. 3, Zurich, Switzerland, 2000; 1546–1555.
33. Rodríguez Pérez M, Herrería Alonso S, Fernández Veiga M, López García C. The persistent congestion problem of FAST-TCP: Analysis and solutions. *European Transactions on Telecommunications* Oct 2010; **21**:504–518.
34. Fu C, Liew S. A remedy for performance degradation of TCP Vegas in asymmetric networks. *IEEE Communications Letters* Jan 2003; **7**(1):42–44.
35. Chan YC, Chan CT, Chen YC. An enhanced congestion avoidance mechanism for TCP Vegas. *IEEE Communications Letters* Jul 2003; **7**(7):343–345.
36. Liu J, Chen F, Wei G. Enhanced TCP Vegas for asymmetric networks. *Wireless Communications, Networking and Mobile Computing*, vol. 2, 2005; 1095–1098.
37. Herrería Alonso S, Rodríguez Pérez M, Suárez González A, Fernández Veiga M, López García C. Improving TCP Vegas fairness in presence of backward traffic. *IEEE Communications Letters* Mar 2007; **11**(3):273–275.
38. Jin C, X Wei D, H Low S. FAST TCP: Motivation, architecture, algorithms, performance. *Technical Report*, Caltech CS Dec 2003.
39. Samios CB, Vernon MK. Modeling the throughput of TCP Vegas. *SIGMETRICS Perform. Eval. Rev.* Jun 2003; **31**(1):71–81, doi:10.1145/885651.781037.
40. Kelly T. Scalable TCP: improving performance in highspeed wide area networks. *Comput. Commun. Review* Apr 2003; **32**(2).
41. Floyd S. HighSpeed TCP. RFC 3649 Dec 1993.
42. Tan L, Zhang W, Yuan C. On parameter tuning for FAST TCP. *IEEE Communications Letters* May 2007; **11**(5):458–460.
43. Mathis M, Floyd S, Romanow A. TCP selective Acknowledgment Options. RFC 2018 Oct 1996.
44. Floyd S, Jacobson V. Random early detection gateways for congestion avoidance. *IEEE/ACM Transactions on Networking* Aug 1993; **1**(4):397–413.
45. Moon SB, Skelly P, Towsley D. Estimation and removal of clock skew from network delay measurements. *Proceedings of the IEEE INFOCOM*, vol. 1, 1999; 227–234.
46. Zhang L, Liu Z, Honghui Xia C. Clock synchronization algorithms for network measurements. *Proceedings of the IEEE INFOCOM*, vol. 1, 2002; 160–169.
47. La RJ, Walrand J, Anantharam V. Issues in TCP Vegas Jan 1999. <http://www.eecs.berkeley.edu/~ananth/1999-2001/Richard/>.
48. Alemu T, Jean-Marie A. Dynamic configuration of RED parameters. *Proceedings of the IEEE GLOBECOM*, 2004; 1600–1604.
49. Athuraliya S, Li VH, Low SH, Yin Q. REM: Active queue management. *IEEE Network* May 2001; **15**(3):48–53.
50. Chan YC, Chan CT, Chen YC, Ho CY. Performance improvement of congestion avoidance mechanism for TCP Vegas. *Proc. ICPADS'04*, 2004; 605–612.
51. Cui T, Andrew L, Zukerman M, Tan L. Improving the fairness of FAST TCP to new flows. *IEEE Communications Letters* May 2006; **10**(5):414–416.
52. Rodríguez Pérez M, Herrería Alonso S, Fernández Veiga M, López García C. Achieving fair network equilibria with delay-based congestion control algorithms. *IEEE Communications Letters* Jul 2008; **12**(7):535–537.
53. Mo J, La RJ, Anantharam V, Walrand J. Analysis and comparison of TCP Reno and Vegas. *Proceedings of the IEEE INFOCOM*, vol. 3, New York, NY, USA, 1999; 1556–1563.
54. Martin J, Nilsson A, Rhee I. Delay-based congestion avoidance for TCP. *IEEE/ACM Transactions on Networking* 2003; **11**(3):356–369.
55. Łukasz Budzisz, Stanojević R, Shorten R, Baker F. A strategy for fair coexistence of loss and delay-based congestion control algorithms. *IEEE Communications Letters* Jul 2009; **13**:555–557.
56. Hayes DA, Armitage G. Improved coexistence and loss tolerance for delay based TCP congestion control. *35th Annual IEEE Conference on Local Computer Networks (LCN 2010)*, Denver, Colorado (USA), 2010; 24–31.